Patent Office
Research and Development
Reports . . . . . No. 10

A FIRST APPROACH TO PATENT SEARCHING PROCEDURES ON STANDARDS ELECTRONIC AUTOMATIC COMPUTER (SEAC)

FOR OFFICIAL DISTRIBUTION

Prepared by

Harold Pfeffer
Herbert R. Koller
Patent Research Specialists
Staff Members, Office of Research and Development
U. S. Patent Office

Ethel C. Marden
Data Processing Analyst
Data Processing Systems Division
National Bureau of Standards

Office of Research and Development

U. S. Patent Office

January 28, 1958

Robert C. Watson
Commissioner of Patents

Sinclair Weeks
Secretary of Commerce

TABLE OF CONTENTS

PREFACE

This report is an amplification of the paper "A First Approach to the Patent Search on a Digital Computer (SEAC)" presented at the 12th National Meeting of the Association for Computing Machinery at Houston, Texas on June 19-20, 1957 and has been available only as an internal report from the authors to the Patent Office and the National Bureau of Standards.

A FIRST APPROACH TO PATENT SEARCHING PROCEDURES ON STANDARDS ELECTRONIC AUTOMATIC COMPUTER (SEAC)

I. INTRODUCTION

In accordance with both a congressional mandate and a recommendation by the Bush Committee,[1] the U.S. Patent Office and the National Bureau of Standards have been engaged in a study of the needs of the U.S. Patent Office in the mechanization of its literature search and retrieval of information.

The necessity for such mechanization is daily growing more pressing. Since the turn of the century the amount of research conducted in all fields of technology has been accelerating; in some fields it is increasing at an exponential rate. There has resulted a consequent increase in all forms of technical literature. In the Patent Office, literature searching forms a large part of the patent examination procedure. This enormous task concerns itself with the scanning of all of the world's technical literature, including both foreign and domestic patents.

The Patent Office has found it more and more difficult to cope with the problem of making an adequate search of this huge (and continually expanding) mass of data. An additional complication for the Patent Office is the statutory requirement that patentability be predicated on novelty, utility, and inventiveness. These considerations must be taken into account, therefore, in the design of a mechanized search system. Only in the last few years has attention been focused on the mechanization of literature searching and information retrieval, although interest in them is now becoming widespread. The use of some sort of high-speed data-processing system, such as an electronic computer, appears to offer the best promise of conducting such searches both rapidly and with the required degree of thoroughness.

For some months the Office of Research and Development of the Patent Office has been determining characteristics of the kinds of searches which it would be desirable to make. Collaboration between the Patent Office and the National Bureau of Standards has resulted in the development of a preliminary search program, which is restricted at this time to the field of chemistry. This program is called HAYSTAQ. A machine code to conduct the search has been written for the NBS electronic computer, SEAC, and trial runs of searches have been made using test data.

The HAYSTAQ system is intended to be used to explore methods of handling literature search problems by high-speed electronic computers and to determine the characteristics and design of equipment best suited for such problems. It is also intended to be used to gather statistics in order to evaluate the efficiency of various features of the system. Finally, it is hoped to evaluate the logic embodied in this program with a view to its possible use as a model or prototype from which improved systems may be derived. A large part of the actual procedure employed consists of "matching" the queries contained in the "Question" (which will be presented by the searcher desiring the information) against information contained in the "Disclosure" (any one of the many documents making up the stored file which is being searched.)

II. THE HAYSTAQ SYSTEM

The HAYSTAQ system includes (1) a data preparation routine for the library making up the complete disclosure file of information to be searched; (2) a data preparation routine for the Question; (3) the search routine, with its included subroutines; and (4) the checkout routine, which evaluates the apparent answers found to questions. Of the four routines, only the third, the search routine, has been written and debugged.

1. The Disclosure Data Preparation Routine

The first-named routine presents a great challenge. It is an encoding, assembly, and data-checking routine to be used for making up a permanent and continually expanding file of the world's technical

literature. The task of encoding the mass of information making up the file, and of ensuring its accuracy, is truly Herculean. Because the file of encoded information is good only so far as it is accurate, ingenious and very fine-sweeping error detection methods must be devised to eliminate the element of human error. Study is now going forward on the preliminary aspects of some of the problems which must be solved before the compilation of such a file.[2]

2. The Question Data Preparation Routine

The second code checks the unedited question data for errors. It generates control words, screening and housekeeping information, extractors for selecting the pertinent parts of the information contained in disclosure data words and sets up variable connectors which will determine the paths which the search routine must follow. It then compiles the edited information and stores it on a magnetic input medium to be used as Question input data for the search routine. These data are set up in the form of a model answer which has the same format as the encoded disclosure data.

3. The Search Routine

In order to describe the third code in the system, which is the search routine, it is necessary to outline the organization of the encoded chemical disclosure file (Figure 1).

Patents or other documents are contained serially in the disclosure file. Each such document is divided into compositions, which represent physical admixtures of materials. Each composition is further subdivided into ingredients or items. An item is made up of a series of descriptor words. The organizational units at the document, composition and item levels are preceded by heading information for identification, screening, and housekeeping functions. Process relationships among compositions are indicated in the composition heading words.[2]

The question is stored in SEAC's high-speed memory and remains there throughout the period of the search. The disclosure file, except for the encoded molecular structure data, is stored on a magnetic tape, called the principal tape. One disclosure composition at a time is read into SEAC's memory and compared against pre-stored Question compositions.

PATENT

COMPOSITION I

1. ITEM A

   DESCRIPTORS

2. ITEM B

   DESCRIPTORS

COMPOSITION II

1. ITEM C

   DESCRIPTORS

2. ITEM D

   DESCRIPTORS

3. ITEM E

   DESCRIPTORS

Figure 1

In conducting the search, the routine progresses through various levels of organization. One question item at a time is compared with each item of the stored disclosure composition. This is done by matching the individual question descriptors making up the question item against the individual disclosure descriptors. If no match is found for any question descriptor, further searching of this item is useless and the search progresses to the next item. If any question item cannot be matched, the search reverts again to the composition level and a new composition is read into SEAC's memory.

Each one of three subroutines may be used in making a detailed search, if called for by the question item. These subroutines are for a chemical descriptor search, an empirical formula match, and a detailed search of the molecular structure. When an apparent answer to a question item is found, a hit word is stored on one of the auxiliary magnetic tape units. This word identifies the question item and the disclosure item which responded to it. The hit word is further marked to indicate whether or not a structure search is required for that item. When apparent answers have been obtained for all question items, the hit words are examined for structure search flags. HAYSTAQ will then make a structure search on those items so marked by the question.[3] The hit words which call for a structure search are subsequently cancelled if a match ~~is not made~~ on the structural part of the question.

The encoded structure data to be used in the structure search subroutine is contained on a separate magnetic tape called the secondary tape. The movement of the two tapes is coordinated by the program so that as the search progresses from one document to another on the principal tape, the corresponding structural information contained on the secondary tape keeps pace with it in order to be available immediately when required. By reading structure information into SEAC only when it is required for the structure search, the needless handling of large amounts of data is avoided.

4. The Checkout Routine

In those cases in which all of the question items in the composition have found answers, the validity of the answers must be ascertained. The block of information comprising a complete set of hit words is therefore read in from tape and processed through the fourth, or checkout, routine. In a search for a mixture of materials, the hit words are analyzed to determine whether the combinatorial relationships among the several items are the ones sought and whether a sufficient number of real answers have been found. This is necessary for several reasons: (1) The procedure in the search is such that one question item may be answered by several disclosure items, or (2) more seriously, one disclosure item may answer several question items. Further, (3), several disclosure items which are wanted only if they are in combination may be disclosed as being in an alternative relationship. In the case of a process question, answers are analyzed to determine whether the various steps of the disclosure process are in the correct sequence. When a complete answer checks out, the identification of the document is printed, and the search continues with the next document.

III. FEATURES OF THE SEARCH

Because of the serial nature of the operation and the huge amount of data which must be examined, the elimination of unprofitable searches is highly desirable. When this can be accomplished by an early rejection of large blocks of data the searching time can be greatly reduced. Two devices which assist this effort are (1) ordering of data and (2) screening.

By ordering the data according to various schemes, HAYSTAQ can determine when continuation of the search at any organizational level is futile. It will then direct the search to the next organizational unit. For example, all descriptors in an item are ordered in an ascending series, according to their first digits, which identify the category of the subject matter (Figure 2). Before comparing descriptors, their first digits are compared. Whenever the first digit of a disclosure descriptor is greater than that of the question, either there is no disclosure descriptor with the same initial digit as the question descriptor being compared, or all of them have already been examined. It therefore becomes useless to look any further in that disclosure item, and the next item is considered.

The screening procedure is also carried out on various levels. Cognizance is taken of certain general requirements of the question and if these are not met by the disclosure data, the unit of disclosure can be rejected immediately. In unprofitable searches, search time can be reduced substantially when rejection of disclosures can be made on the document level. For example, if the question requires a particular process, the document is examined to determine whether <u>any</u> process is included. If not, the entire document is eliminated from the search. If a process is detected, the heading words of the document are examined to see whether the process contains at least as many compositions as are required by the question. If not, the document is rejected. Failure to pass any screen results in rejection of the unit of disclosure being examined. If all screens are passed on the document level, the search continues at more intensive levels of first the composition, then the item.

A great part of the strength of the HAYSTAQ program lies in its ability to handle diverse relationships among chemical compounds per se and among their structural elements. (See Appendix C) Already mentioned is the ability to search for compounds, compositions and chemical processes. There are also included some more sophisticated types of searches which are not readily available to the examiner in the manual system in use today. One of these is the ability to make alternative searches, in which one or more items are stated to be in an "and" relationship with respect to a group of items which are in an "or" relationship: e.g., A <u>and</u> B, <u>and either</u> C <u>or</u> D. There is also provided the ability to search for a "teaching"* of equivalence in a disclosure document; that is to say, that the disclosure document itself provides the statement that one compound may be substituted for another in a given set of circumstances.

---

* "teaching" as used in this paper is a statement made in a disclosure document.

EXAMPLES OF DESCRIPTOR TYPES

10000004852	-	Index Number
2043C05E901	-	Empirical Formula
3----------	-	Chemical
4----------	-	Botanical
5----------	-	Zoological
6----------	-	Anatomical
7----------	-	Processes
8----------	-	Miscellaneous

Figure 2

An additional feature is the ability to formulate a question with respect to negative teachings. In this case, any document which would provide an acceptable response to the question must contain the statement that a certain thing is not present under given circumstances. A corollary to this situation is the ability to ask that a specified thing be absent. For this search, a responding document may "teach" that the specified thing must be absent, or it may fail to mention it at all, the sole requirement being that there is no teaching of its presence. The inclusion of these negative concepts along with the positive concepts in both question and disclosure results in fifteen different variations in the possible search paths which HAYSTAQ might follow. (Figure 3. See also Appendix A).

QUESTION		DISCLOSURE
Is there an item which has		An item is disclosed which has
1	All descriptors positive	1
2	Some descriptors positive - others negative	2
3	All descriptors negative	3
4	Some descriptors positive - others absent	-
5	All descriptors absent	-

Figure 3

hese are based on five possible question item types which might be sked against three possible disclosure item types. Coded information efining the different question and disclosure item types is stored in the em heading words. In order to pre-set the particular path to be ollowed to the exclusion of the other fourteen paths, numerical values re assigned to the disclosure types and other numerical values to the uestion types. The addition of the two numbers results in a unique umber for each combination of question-and-disclosure types, and its alue is used to determine the address of the first instruction along the esired path (Figure 4).

Question Item Type	A= 1st digit of Q	F= 200+ A+B	B= 1st digit of D	Disclosure Item Type
All Q descriptors positive	0	200	0	All D descriptors positive*
	0	201	1	Some D descriptors positive, others negative △
	0	202	2	All D descriptors negative**
Some Q descriptors positive, others negative	3	203	0	*
	3	204	1	△
	3	205	2	**
All Q descriptors negative	6	206	0	*
	6	207	1	△
	6	208	2	**
Some Q descriptors positive, others absent	9	209	0	*
	9	210	1	△
	9	211	2	**
All Q descriptors absent	12	212	0	*
	12	213	1	△
	12	214	2	**

Figure 4. Schematic Derivation of Appropriate Address for Item Searching Path, Designated as "F"

The empirical formula subroutine (Appendix B) examines disclosure data stored in the following way. Seven binary digits are used to define uniquely the element and six binary digits denote the number of times it occurs (Figure 5). Thus, thirteen binary digits are required to describe the presence of each different element of a compound. As many of the thirteen-bit combinations as required are stored in succession, and when one SEAC word (of 44 binary bits) is filled, another is begun. When the last element has been recorded, a minus sign is affixed to the last word of the descriptor so that the subroutine can determine when it has finished examining a single descriptor.

IV. MODIFICATIONS OF SEAC

SEAC's limitations made its handling of HAYSTAQ cumbersome in certain respects. Steps were therefore taken to eliminate some of the limitations. One of the most serious of these was size of memory. When operating in the three-address mode, SEAC had 1536 words of high-speed memory: 1024 words of electrostatic and 512 words of mercury delay-line memory. There is now being installed another 512 words of mercury delay-line memory, which will raise the total memory capacity to 2048 words.

SEAC's magnetic input and output media comprise several magnetic wire units, which employ adapted Pierce wire recorder magnetic cartridges. In addition, there are several special magnetic-tape units from which the tapes are usually not removed, which are used principally for temporary or intermediate storage of results or as an auxiliary memory. Access to them, for either input or output, is possible only through SEAC itself. Neither of these media can handle a fraction of the immense disclosure file which the search program must examine. To meet the need for a large capacity high-speed magnetic input medium, eight AMPEX tape units are being installed. The disclosure file will be stored on six-channel tapes, which will be approximately 2,000 feet long, with a packing density of 200 bits to the inch and a speed of 40 to 60 inches per second. SEAC will be capable of sensing two sizes of gaps, thus permitting differentiation between two kinds of variable length records. Since the unit of input is a composition the smaller gaps could be used to separate compositions. The larger gaps could be placed between documents, since many times

	Identification of Empirical Formula Word	*	Hydrogen; Atomic Number = 1	Three occurrences	Carbon; Atomic Number = 6	Twelve Occurrences	Oxygen; Atomic Number = 8	Four Occurrences	**
Hexa-Decimal Code	2	0	01	03	06	0C	08	04	-
Binary	0010	0	000 0001	00 0011	000 0110	00 1100	0001000	00 0100	1
No. of Bits	4	1	7	6	7	6	7	6	1

* In D, this bit = 0; in Q, it is zero for "at least" or 1 for an exact match.

** This bit is in the sign position. Minus, or 1, indicates last word of empirical formula; plus or 0 denotes incomplete.

Figure 5. Format of Empirical Formula Word. Machine code is illustrated on line marked "Binary". The empirical formula shown is $H_3C_{12}0_4$.

rejection is made at the document level. When one unit of information is rejected, SEAC can be instructed to advance the tape to tne next gap of the desired size. If the instruction to the tape unit is to proceed to the next long gap (that is, to the next document), then while that tape is advancing, examination of the disclosure data stored on the next available tape can begin. This procedure permits the search to progress from one tape to another after each fruitless attempt until the cycle has been completed and the first tape is reached again. It is anticipated that most of the documents examined will be discarded either as a result of the early coarse screens or at some later stage in the search.

Along with the changes in memory capacity and input media, two new orders were added to SEAC's repertoire for greater ease in the handling of HAYSTAQ. These are: a shift order, for either right or left shift, and an equality comparison order. Before these additions to SEAC's command capabilities, a shift was accomplished by multiplication or division by the required binary numbers, and equality could be determined only by making two comparisons, using either SEAC's algebraic or absolute value comparison orders.

V. CONCLUSION

Several arbitrary limitations have been imposed on the HAYSTAQ system in order to accomplish some of the aims of the investigation at an early date. For example, the system is now limited to searches in the field of chemistry. While the general concepts on which this system is based are believed applicable to subject matter in fields other than chemistry, no attempt has as yet been made to apply the system to searches in other fields, such as the mechanical or electrical fields. The system permits many varied types of questions, some of which present numbers of complex variations. Limitations have been imposed upon some of these because of the impracticability of exploring all of the possibilities at the present time.

An inherent limitation of the entire system, which will become more apparent as the encoded library in its growth approaches the anticipated order of magnitude, is the serial nature of inspection of the data contained in the documents. The search progresses, for each question, from one document to another until the last document in the gargantuan file has been examined. The mechanized searching system described here is designed to be used with a file of data which will ultimately include the encoded information contained in several million technological documents.

It is likely that present-day machine concepts may be unsuited for large-scale literature searching operations. This idea suggests that future research might well be conducted in the design of machine systems which could search a static disclosure file in a parallel manner. A parallel method of approach to the searching problem appears to offer the ideal solution for this type of problem. It may be that machines will have to be designed and built which are only remotely related to the computers known today. One such machine might possibly be a giant "comparator" which could simultaneously compare a multitude of stored disclosures against a Question. Some combination of parallel and serial approach might well offer an acceptable transitional solution.

The problem of mechanization is universally faced by all who have a need for literature searching and information retrieval. It is hoped that this beginning may serve as a stimulus toward the development of more sophisticated routines, and to research leading to the development of better machines for handling this type of problem. It is hoped, also, that attacks will begin on a large front on the problem of creating an unambiguous, technical language suitable for making the world's technical literature available to the world's researchers.

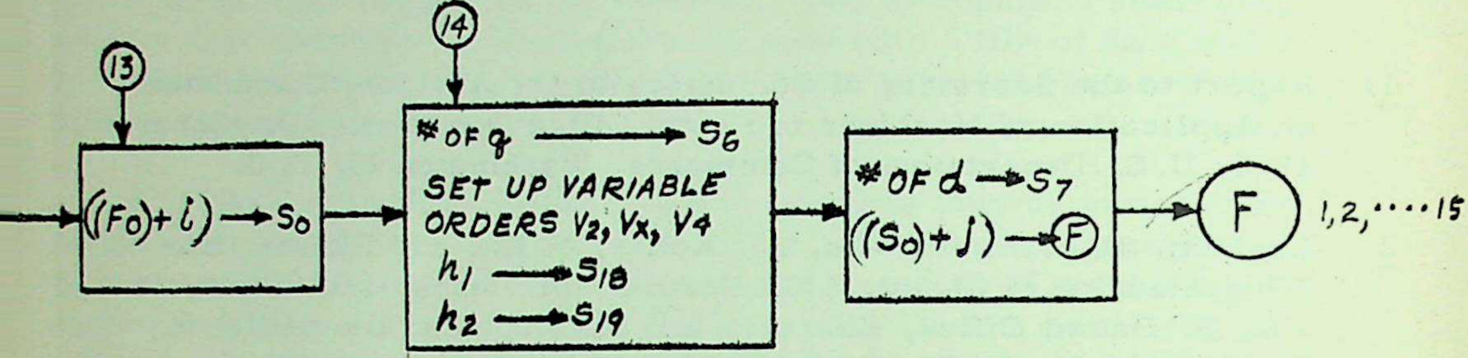
BIBLIOGRAPHY

1/ Report to the Secretary of Commerce by the Advisory Committee on Application of Machines to Patent Office Operations; December 22, 1954, U. S. Department of Commerce, Washington 25, D. C.

2/ Lanham, B. E., Leibowitz, J., Koller, H. R., and Pfeffer, H.; "Organization of Chemical Disclosures for Mechanized Retrieval" - U. S. Patent Office, Research and Development Report No. 5., Washington 25, D. C., June 14, 1957.

3/ Ray, L. C., and Kirsch, R. A., - "Finding Chemical Records by Digital Computers" - SCIENCE, Vol. 126, page 814, October 25, 1957.

APPENDIX A



Detail of Flow Chart Selecting Item Subroutine

i is the number representing the question type

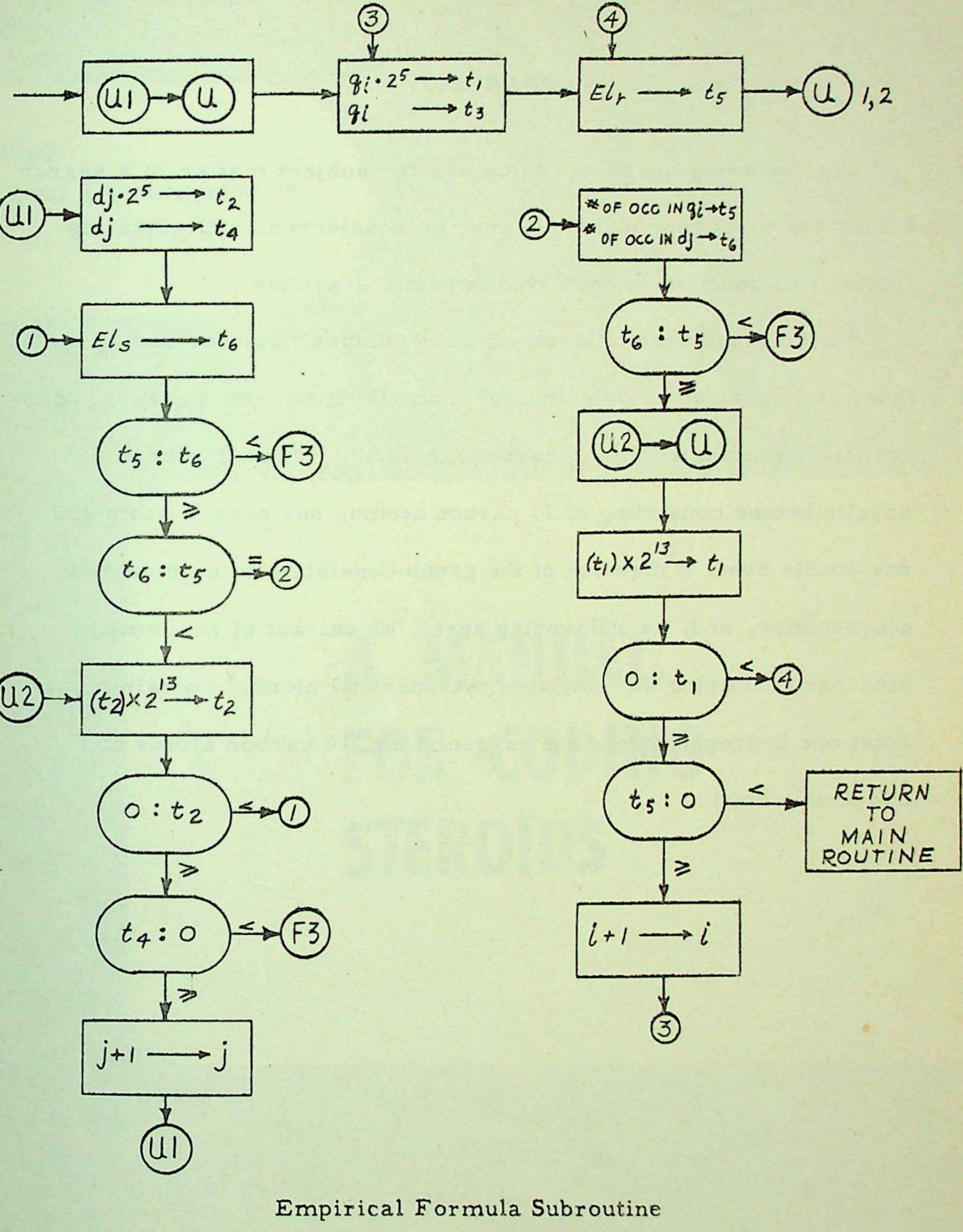
j " " " " " disclosure type

h_1 and h_2 contain housekeeping information for a new disclosure item

F_0 is the base address for the 15 item subroutines

In box 13 the question type plus the base address is held constant in S_0, to be used with successive disclosure items. Control enters this part of the routine through box 14 for a new disclosure item, and through box 13 for a new question item.

APPENDIX B



Empirical Formula Subroutine

q_i = question descriptor word (i = 1, 2, ... n)

d_j = disclosure " " (j = 1, 2, ... m)

El_r = any element in question empirical formula

El_s = " " " disclosure " "

APPENDIX C

The following question, which was the subject matter of a search during the debugging operation, may be considered as illustrating typical relationships encountered in patent searches:

Find a disclosure of a non-aqueous composition, for use against fever and ague, comprising boneset, an alkylated cyclohexenyl hydrocarbon containing at least 7 carbon atoms and one double bond, an acyclic ketone consisting of 11 carbon atoms, one oxygen atom and one double bond, a member of the group consisting of quinine and cinchonidine, and, as a flavoring agent, an extract of the family pinaceae containing an alkylated cyclohexenyl alcohol containing at least one hydrogen atom, one oxygen atom, 10 carbon atoms and one double bond.